MATRIX COMPUTATION ENGINE

[0001] This application is a continuation of U.S. patent application Ser. No. 15/800,342, filed on Nov. 1, 2017. The above application is incorporated herein by reference in its entirety.

BACKGROUND

Technical Field

[0002] Embodiments described herein are related to circuitry to perform matrix operations in processor-based systems.

Description of the Related Art

[0003] A variety of workloads being performed in modern computing systems rely on massive amounts of matrix multiplications. For example, certain long short term memory (LSTM) learning algorithms are used in a variety of contexts such as language detection, card readers, natural language processing, and handwriting processing, among other things. LSTM processing includes numerous matrix multiplications. The matrix multiplications may be small integers, for example, but very large numbers of them. The performance of such operations on a general purpose central processing unit (CPU), even a CPU with vector instructions, is very low; while the power consumption is very high. Low performance, high power workloads are problematic for any computing system, but are especially problematic for battery-powered systems.

SUMMARY

[0004] In an embodiment, a matrix computation engine is configured to perform matrix computations (e.g. matrix multiplications). The matrix computation engine may perform numerous matrix computations in parallel, in an embodiment. More particularly, the matrix computation engine may be configured to perform numerous multiplication operations in parallel on input matrix elements, generating resulting matrix elements. In an embodiment, the matrix computation engine may be configured to accumulate results in a result memory, performing multiply-accumulate operations for each matrix element of each matrix. The matrix computation engine may be both high performance and power efficient, in an embodiment, as compared to a general purpose processor (even one with vector instructions), for example.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] The following detailed description makes reference to the accompanying drawings, which are now briefly described.

[0006] FIG. **1** is a block diagram of one embodiment of a processor, a matrix computation engine, and a lower level cache.

[0007] FIG. **2** is a block diagram illustrating one embodiment of X, Y, and Z memories and a multiply-accumulate (MAC) circuit for the matrix computation engine shown in FIG. **1**.

[0008] FIG. **3** is a block diagram illustrating one of MACs generating result matrix elements for one embodiment.

[0009] FIG. **4** is a block diagram illustrating matrix element value remapping for one embodiment.

[0010] FIG. **5** is table of instructions which may be used for one embodiment of the processor and matrix computation engine.

[0011] FIG. **6** is a block diagram of one embodiment of a system.

[0012] FIG. **7** is a block diagram of one embodiment of a computer accessible storage medium.

[0013] While embodiments described in this disclosure may be susceptible to various modifications and alternative forms, specific embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that the drawings and detailed description thereto are not intended to limit the embodiments to the particular form disclosed, but on the contrary, the intention is to cover all modifications, equivalents and alternatives falling within the spirit and scope of the appended claims. The headings used herein are for organizational purposes only and are not meant to be used to limit the scope of the description. As used throughout this application, the word "may" is used in a permissive sense (i.e., meaning having the potential to), rather than the mandatory sense (i.e., meaning must). Similarly, the words "include", "including", and "includes" mean including, but not limited to.

[0014] Within this disclosure, different entities (which may variously be referred to as "units," "circuits," other components, etc.) may be described or claimed as "configured" to perform one or more tasks or operations. This formulation—[entity] configured to [perform one or more tasks]—is used herein to refer to structure (i.e., something physical, such as an electronic circuit). More specifically, this formulation is used to indicate that this structure is arranged to perform the one or more tasks during operation. A structure can be said to be "configured to" perform some task even if the structure is not currently being operated. A "clock circuit configured to generate an output clock signal" is intended to cover, for example, a circuit that performs this function during operation, even if the circuit in question is not currently being used (e.g., power is not connected to it). Thus, an entity described or recited as "configured to" perform some task refers to something physical, such as a device, circuit, memory storing program instructions executable to implement the task, etc. This phrase is not used herein to refer to something intangible. In general, the circuitry that forms the structure corresponding to "configured to" may include hardware circuits. The hardware circuits may include any combination of combinatorial logic circuitry, clocked storage devices such as flops, registers, latches, etc., finite state machines, memory such as static random access memory or embedded dynamic random access memory, custom designed circuitry, analog circuitry, programmable logic arrays, etc. Similarly, various units/circuits/components may be described as performing a task or tasks, for convenience in the description. Such descriptions should be interpreted as including the phrase "configured to."

[0015] The term "configured to" is not intended to mean "configurable to." An unprogrammed FPGA, for example, would not be considered to be "configured to" perform some specific function, although it may be "configurable to" perform that function. After appropriate programming, the FPGA may then be configured to perform that function.

[0016] Reciting in the appended claims a unit/circuit/component or other structure that is configured to perform